

Define.xml: Good Practices and Stylesheets

Jozef Aerts
XML4Pharma



What is define.xml ?

- **metadata** about your SDTM/SEND/ADaM submission
- an **XML** file
- separates **content** from presentation
 - presentation by a stylesheet

XML basics

- XML is **case sensitive**
 - so needs to be the content of your define.xml
- uses Unicode
 - but define.xml samples still use ISO-8859-1 (Latin-1 encoding)
 - recommendation: use **UTF-8** encoding
- Define.xml is an **ODM extension**
 - everything that is allowed in ODM is allowed in define.xml
 - but does not always make sense ...

Dataset and Variable ordering

- 1 Dataset = 1 `ItemGroupDef`
- Datasets are expected in the order:
 - trial design
 - special purpose
 - interventions
 - events
 - findings
 - relationships
- and **alphabetically** within each class

Dataset and Variable ordering

- Variables ("ItemRef") should come in same order as in the datasets
- **OrderNumber** is superfluous
 - usually **ignored** by the stylesheet anyway

The "IsReferenceData" attribute

- "Yes" for datasets that are for reference, e.g. trial design
- "No" for datasets with "actual" metadata

```
- <ItemGroupDef OID="TA" Name="TA" SASDatasetName="TA" Repeating="No" IsReferenceData="Yes" Purpose="Tabulation"
  def:Label="Trial Arms" def:Structure="One record per planned Element per Arm" def:DomainKeys="STUDYID, ARMCD,
  TAETORD" def:Class="Trial Design" def:ArchiveLocationID="Location.TA">
  <!-- ***** -->
  <!-- Each variable is listed here for this ItemGroupDef (domain) -->
  <!-- ***** -->
  <ItemRef ItemOID="STUDYID" OrderNumber="1" Mandatory="Yes" Role="IDENTIFIER" RoleCodeListOID="RoleCodeList" />
  <ItemRef ItemOID="DOMAIN" OrderNumber="2" Mandatory="Yes" Role="IDENTIFIER" RoleCodeListOID="RoleCodeList" />
  <ItemRef ItemOID="TA.ARMCD" OrderNumber="3" Mandatory="Yes" Role="TOPIC" RoleCodeListOID="RoleCodeList" />
  <ItemRef ItemOID="TA.ARM" OrderNumber="4" Mandatory="Yes" Role="SYNONYM QUALIFIER" />
```

The "Repeating" attribute (ItemGroupDef)

- "Yes" for domains with the potential of having more than one record per subject
- "No" for domains restricted to have one record per subject (e.g. DM, SUPPDM)
- "No" for trial design domains

```
- <ItemGroupDef OID="DM" Name="DM" SASDatasetName="DM" Repeating="No"
  IsReferenceData="No" Purpose="Tabulation" def:Label="Demographics" def:Structure="One
  record per subject" def:DomainKeys="STUDYID, USUBJID" def:Class="Special Purpose"
  def:ArchiveLocationID="Location.DM">
  <!-- ***** -->
  <!-- Each variable is listed here for this ItemGroupDef (domain) -->
  <!-- ***** -->
  <ItemRef ItemOID="STUDYID" OrderNumber="1" Mandatory="Yes" Role="IDENTIFIER"
  RoleCodeListOID="RoleCodeList" />
```

The "Mandatory" attribute (ItemRef)

- SDTM "required" => "Yes"
- SDTM "permissible" => "No"
- *SDTM "expected" => "No"*

The "Role" attribute (on ItemRef)

- Identifier
- Topic
- Timing
- Synonym Qualifier
- Grouping Qualifier
- Variable Qualifier
- Record Qualifier
- Result Qualifier
- Rule

**Can usually be copied from the SDTM/ADaM/SEND Implementation Guide
but you need to find out yourself for non-standard variables (for SUPP--)**

Do not forget to add "CodeList" for the roles
and reference it by "RoleCodeListOID"

ItemDef attributes: Datatype, Length, SignificantDigits

- SAS char => text
 - except for ISO-8601:
 - date, time, datetime
 - --DUR => text
- SAS numeric
 - only integers => integer
 - floats (+integers) => float
- mixed => ... be smart ...

ItemDef attributes:

Datatype, Length, SignificantDigits

- **Length**: only for integer, float, text
 - is a "character" length, not a SAS length
 - e.g. "123.456" => Length=7
- **SignificantDigits**: only for float
 - number of characters after the decimal point
 - e.g. "123.456" => SignificantDigits=3

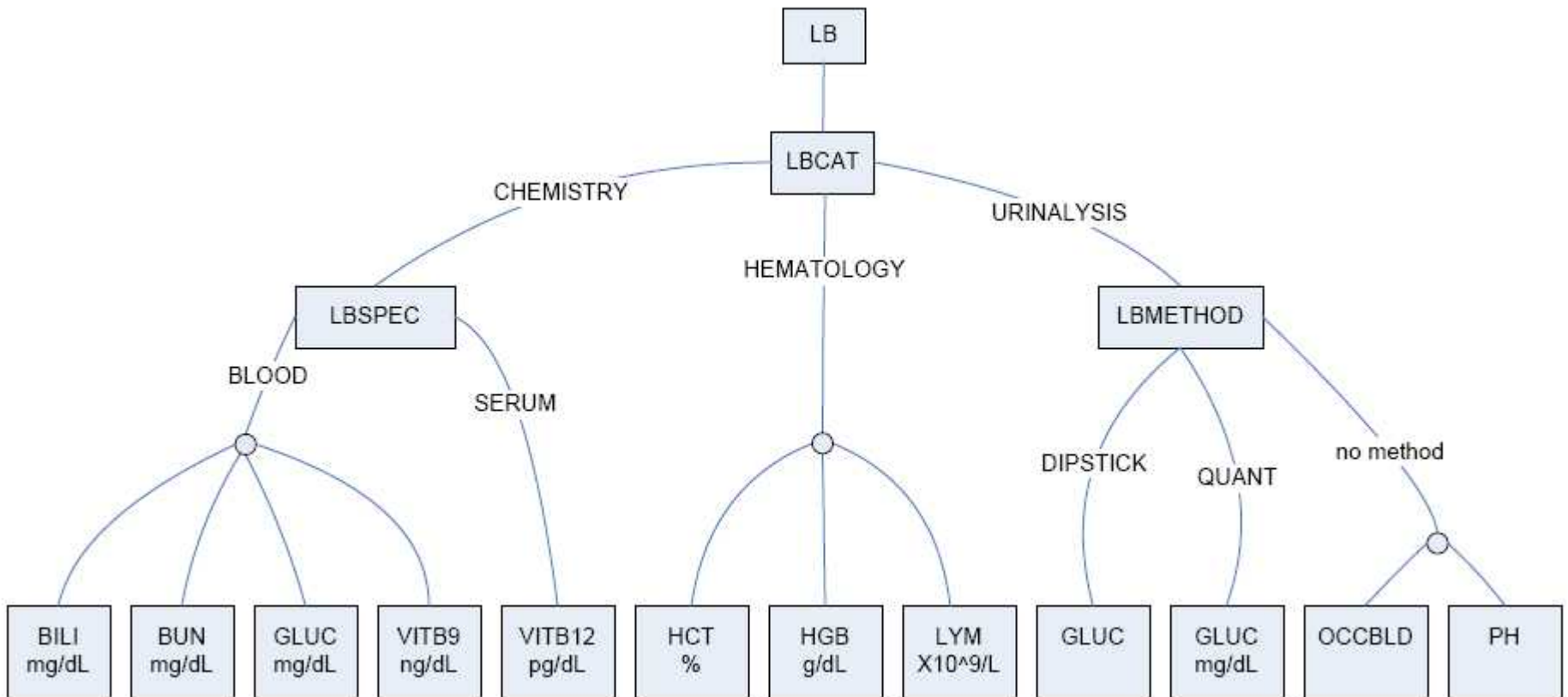
ValueLists

- listens possible values for --TESTCD, QNAM
- each value has a datatype, length, ...
 - e.g. VSTESTCD:
 - **WEIGHT**: float
 - **HEIGHT**: integer
 - **FRAME**: text with controlled terminology
- ValueLists are not applicable to --STRESC, --STRESU

ValueLists

- Nested valuelists often for LB:
 - LBCAT, LBSPEC, LBMETHOD
 - can become pretty complicated
 - recommendation: use "**MeasurementUnitRef**" at deepest "ItemDef" level
 - alternative: splitted datasets
- For QS: often better to split the dataset

ValueList example from the "Metadata Submission Guidelines"



Courtesy: R.Lewis, Octagon Research

CodeLists

- list controlled terminology for variables
- use CDISC-CT as much as possible
- some CDISC CT is **extensible**, some not
- **CASE SENSITIVE !!!**
 - usually **uppercase** for CDISC-CT
 - lower/mixed case for external CT
- code and decode are identical

```
- <CodeList OID="ACN" Name="ACN" DataType="text">  
- <CodeListItem CodedValue="DOSE NOT CHANGED">  
  - <Decode>  
    <TranslatedText>DOSE NOT CHANGED</TranslatedText>  
  </Decode>  
</CodeListItem>
```

CodeLists

- xml:lang ignored by stylesheet
 - can be either
 - xml:lang="en"
 - or **absent**
- External CT: use "ExternalCodeList"
 - don't forget to add **Version**

```
- <CodeList OID="AEDICT_F" Name="ADVERSE EVENT DICTIONARY" DataType="text">
  <ExternalCodeList Dictionary="MEDDRA" Version="8.0" />
</CodeList>
- <CodeList OID="DRUGDICT_F" Name="DRUG DICTIONARY" DataType="text">
  <ExternalCodeList Dictionary="WHODRUG" Version="200204" />
</CodeList>
```


Generation of define.xml

- partial define.xml can be generated from the SAS XPT datasets, e.g. OpenCDISC
- partial define.xml needs to be extended
 - using an XML editor (not XML-Spy)
 - Define.xml White Paper
 - using the "define.xml designer"
- better is to keep define.xml in sync during mapping
 - => SDTM-ETL

Why a stylesheet?

- define.xml = content, not presentation
- stylesheet allows presentation as HTML
- uses XSLT technology
- FDA requires you to provide a stylesheet

If the FDA were wise

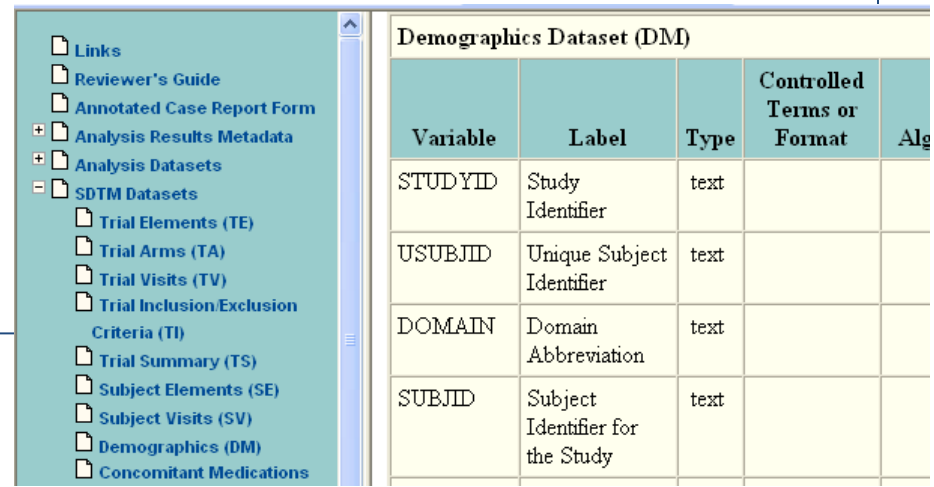
- They would **FORBID** you to provide a stylesheet
- They would have their own (set of) stylesheets
- They would have their OWN views on YOUR metadata
 - one set of data, multiple views
- However ...

FDA and XML technology

- There is no XML / XSLT knowledge at the FDA
- The FDA is not able to create stylesheets
- Opportunity for the sponsor?
 - to provide the FDA a sponsor-centric view of the metadata

Stylesheets – design decisions

- which HTML output ?
- for which browsers / versions ?
- use of JavaScript ?
- table of contents ?
 - as collapsible/extensible tree view?



The screenshot shows a web application interface. On the left is a tree view with a teal background, listing various data categories. On the right is a table titled "Demographics Dataset (DM)" with a yellow background. The table has five columns: Variable, Label, Type, Controlled Terms or Format, and Alg.

| Variable | Label | Type | Controlled Terms or Format | Alg |
|----------|----------------------------------|------|----------------------------|-----|
| STUDYID | Study Identifier | text | | |
| USUBJID | Unique Subject Identifier | text | | |
| DOMAIN | Domain Abbreviation | text | | |
| SUBJID | Subject Identifier for the Study | text | | |

Browsers and versions

- What is the FDA using ?
 - talk to them
- Most define.xsl stylesheets
 - work well in IE7
 - disastrous in IE8 (TOC)
 - IE8 can run in IE7-compatibility mode
 - problem is JavaScript
 - alternative: don't use JavaScript

References to annotated CRF or supplementalDoc

- PDF hyperlinks
- work / don't work depending on PDF version and Adobe Viewer version
 - "CRF page ..." does not always ...
- Recommendation: only use "named destinations"

Some misconceptions

- Attribute order
- ItemDef order
- Codelist and ImputationMethod order
- OID attribute is not fixed by the standard
 - however "Name" is fixed

Validation

- Read the "XML Schema Validation for Define.xml White Paper"
 - see: <http://www.cdisc.org/define-xml>
- Use tools
 - OpenCDISC Validator
 - Define.xml Checker
- Let someone play "the reviewer"

Conclusions

- Define.xml *is* XML
 - so try to think in XML ...
 - RTFMs
 - Read the upcoming
"Metadata Submission Guidelines"
 - use the CDISC Discussion Forum
- If problems with stylesheets => outsource